Morphosemantems Decomposition and Semantic Representation to allow Fast and Efficient Natural Language Recognition

Christian Lovis [i], MD, Robert Baud [ii], PhD,
Pierre-André Michel [ii], Jean-Raoul Scherrer [ii], MD

Medicine Department [i] and Medical Informatics Centre [ii]
Geneva University Hospital

Background. Most natural language processing systems (NLP) are for now limited due to the limited amount of knowledge in their dictionaries. This limitation can be observed in two different aspects, the lexical coverage on one side, and the semantic associated to the lexical information on the other side. Moreover, the medical being especially like to invent new words and NLP systems will have to cope with them. Automatic extraction of knowledge from large corpus of texts is an essential step toward linguistic knowledge acquisition in the medical domain. The requirement to morphologically recognise any word of a sentence before being able to analyse the meaning of the whole expression and the need to have at least one concept attached with any morphological unit is a mandatory step to achieve before that any NLP system can have a practicable use. This step is concomitant with the computational handling of inflectional morphology, as treated by Koskenniemi[1] and supports the idea of dealing with word segmentation methodology for NLP systems.

Methodology. To create the basic entries of our dictionaries, we used normalised source of terms available through multilingual international classifications like ICD. The availability of these term sources in many languages is of priceless help to built the lexical sources and define precisely the conceptual coverage. The linguistic and semantic knowledge has been extracted in five steps : *a) Multilingual inter-classification* pattern mapping has been used to increase the overall language knowledge coverage. This method starts from any pair of expressions sharing the same code in two languages and build an "interlingua" co-occurrence matrix. Taking one word in an expression of the first language and all the words in the expression of same code in the other language gives a number of pairs. Each word of a pair has a pattern of co-occurrence to be compared with the pattern of the other word in its pair. A score of the number of exact match is determined. *b) Normalisation* of words in their basic forms with lexical information, inflexions and typographical variants. *c) Morphosemantic decomposition* of all words in their smallest meaningful units with strong human validation. There are many different elements that may compose a word, including other valid nouns (*congresswoman*) or other lexical elements (*restless*), prefixes (*hyper-, pseudo-*), suffixes (*-therapy, -ectomy*) or inflections (like *-s* or *-ies* for plural)[2]. A given unit may have kept its original meaning (like *-andro* = male) or may have undergone a semantic sludge. To illustrate this situation, we can take *-meno* that means, nowadays, *menstruation* as used in *menopause* or *dysmenorrhea,* instead of *mên = moon* in Greek. *d) Aggregation* of basic forms in semantically similar or homogenous groups. The segmented elements, that represent a kind of atomic concept, are grouped following semantic similarities. These similarities include lexical variants (*lateral - latero*); nouns or adjectives (*cortex - cortical*); synonyms or analogous semantic contents (*heart - myocarde*); abbreviations (*mb - member - membrane*); logos (*AMI - acute + myocardial + infarction*). *e) Conceptual link* with the GALEN model when possible.

Conclusion. The use of powerful segmentation processing linked with semantic representation have prove to be highly efficient for the implementation of a natural language encoder. Such a representation with a tree-based model of runtime dictionaries allows a very fast morphological analysis with semantic recognition to be carried out on a PC based computer. The technique is presently used in several Hospitals and supports the WHO's International Classification of Diseases version 9 and 10 in French, English and German.

References

1. K. Koskenniemi. Two-level model for morphological analysis. PhD Thesis. University of Helsinki, 1983

2. S. Wolff. The Use of Morphosemantic Regularities in the Medical Vocabulary for Automatic Lexical Coding. Meth Inform Med, 23: 195-203, 1984

1091—8280/97/$5.00